Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

EMAIL SPAM DETECTION USING MACHINE LEARNING ALGORITHMS

Mr.L .Sreedhar Professor, CSE department, Raghu Engineering College, Dakamarri, Visakhapatnam

B.Kavya, H.Sai Kiran, Ch.Vinay Bhaskar fourth year CSE students of Raghu Engineering College, Dakamarri, Visakhapatnam

[1] sreedhar.lolla@raghuenggcollege.in, [2]19981a0529@raghuenggcollege.in,
[3]20985a0508@raghuenggcollege.in ,[4] 19981a0533@raghuenggcollege.in

Abstract

Spam emails have consistently been a problem for computer security. They are extremely risky for computers and networks and expensive economically. The importance of email communication has increased over time despite the growth of social networks and other Internet-based information exchange platforms, making it urgently necessary for better spam filters. Even though numerous spam filters have been introduced to help prevent these spam emails from reaching a user's mailbox, there is a paucity of research focusing on text modifications. Due to its efficiency and simplicity, Naive Bayes is currently among the most popular methods for categorizing spam.We will detect spam mails using Naive Bayes in comparison to logistic regression accuracy, which is 96.77%, our Python technique increases Naive Bayes' accuracy by over 97.30%.

Keywords

Machine Learning, Naive Bayes, Logistic Regression,Classification

1. Introduction

Spam is a term for emails that are intended to either randomly overwhelm the inbox or to manipulate the recipient. It also goes by the name of junk mail and overflows Internet users' inboxes. Today's spam emails come in a wide range of forms, including advertisements, business promotions, dubious goods, and some offensive services [1]. As a result, it might be challenging to determine whether an email is a spam or not.

Usenet, also known as User Network, is an email service that disseminates group emails or talks that are primarily instructive but do fill up the user's mailbox. These emails or talks are directed toward a certain group of people linked with a particular service or product. The information that travels through the Internet is referred to as Net news, and a collection of this information intended to spread messages on a certain subject is referred to as a "newsgroup." Spammers specifically target readers of such news from various newsgroups. These newsgroups are used by spammers to advertise various irrelevant adverts or irrelevant messages. Usenet spam undermines the value of newsgroups for users by boosting unrelated posts [2]

Electronic mail, or emails, have grown in popularity as communication becomes more digital; in 2016, an estimated 2.3 million people utilized email. Daily email sending and receiving reached 205 billion in 2015, and it is anticipated that this number would increase to over 246 billion by 2019 (up 3% annually). Unfortunately, because existing spam detection techniques lack an accurate spam classifier, the boom in emails has also resulted in an unprecedented rise in the amount of illegitimate mail, or spam - 49.7% of emails sent are spam [3]. Spam is a concern not just because it frequently carries the virus, but also because spam emails consume a lot of computing, storage, and

network resources. The commercial sphere also has a significant interest in detecting spam.

The next part of the paper is followed by a literature review in section 2, the methodology in section 3, results in section 4 and it contains conclusion and future enhancements in the last section.

2. Literature Review

Karim et al, [4] discussed a focused literature review of machine learning and Artificial Intelligence (AI) methods for email spam detection.T. Kumar and Agarwal. The "image and textual data set for the e-mail spam detection with the use of various methods" was used by Harisinghaney et al. (2014) and Mohamad & Selamat (2015).

Harisinghaney et al. (2014) [5] used methods of KNN algorithm, Naive Bayes, and Reverse DBSCAN algorithm with experimentation on data set. OCR library" is used for text recognition unfortunately this didn't work well.

PSOs can use the stochastic distribution's property to first identify a local search solution, after which each particle shares its own to produce a global solution. Based on the keywords present in the email textual data, NB with probability distribution property determines the potential class for the email content from the spam class or non-spam [6].

The Support Vector Mechanism algorithm is used in this study to detect spam emails. descriptions of the data set used in this study as given on the Spam Assassin website. SVM is also regarded as a crucial kernel method, which is one of the most crucial areasin the theory of machine learning [7].

Both machine learning and non-machine learning techniques were applied in this paper. Methods of machine learning include neural networks and support vector machines. approaches that are not based on machine learning, such as extensive keyword searches and word white- and black-listing. The sets that result from this process are then used as training sets and classification sets. Depending on how much of a spam score each emailreceives, the emails are categorized [8].

3. Methodology

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Fig 1.Flow Chart

The above flow chart represents the flow of spam detection. We first take the dataset then it is passed through the leetspeak converter which converts all the leetspeak characters into normal text.Finally, the dataset is fed to this model which predicts the output of whether the given text is spam or ham. The dataset is sent to the leetspeak converter where all the leetspeak words are converted to corresponding words. Then it is sent to the data pre-processing where all the stop and unwanted words are removed. The data is split into test data and train data. Each of these split data is sent to the model to evaluate their efficiency for the test and train data. Finally, the output is predicted for the test data.

In pre-processing, we remove all the unwanted words and punctuations such as commas,full stops, extra spaces, and other repeated words. This will help to reduce the data to be processed. This helps to process the data much faster than before due to the small amount of data.

We split the data into train and test data for the Naive Bayes model before converting the data to the desired matrix format. Count Vectorizer() will be used to accomplish this Here, there are two steps to think about: Our training data (X train) must first be fitted into Count Vectorizer() before the matrix can be returned. Second, to return the matrix, we must transform our testing data (X test). Fit the training data into the Multinomial classifier using fit after importing it (). Call it "classifier" on your classifier. We can now use predict to make some predictions on the test data stored in the "X test" after our algorithm has been trained using the training data set ().

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Logistic regression is a technique where it calculate the probability of two events.To calculate the likelihood of an event occurring, it is used to fit the event into the logistic function. Spam is denoted by 1 and ham by 0. Here, the data is divided into train and test data before being transformed into the desired matrix format. Count Vectorizer will be used to accomplish this (). Here, there are two steps to think about: Our training data (message train) must first be fitted into Count Vectorizer() before it can return the matrix. Second, to return the matrix, we must transform our testing data (message test).

Fit the training set of data into the model using fit and import the Logistic Regression algorithm (). Give your model the name "spam model." We can now use predict to make some predictions on the test data stored in the "message test" after our algorithm has been trained using the training data set ().

4. Results

The output screenshots shown in figs. 2, 3,4,5,6,7 below are displayed to show the spam data set where the data set contains 2 columns and 5572 rows in which it is noted as Category and Message where Category is notation number and Message is all about Spam and Ham mails.

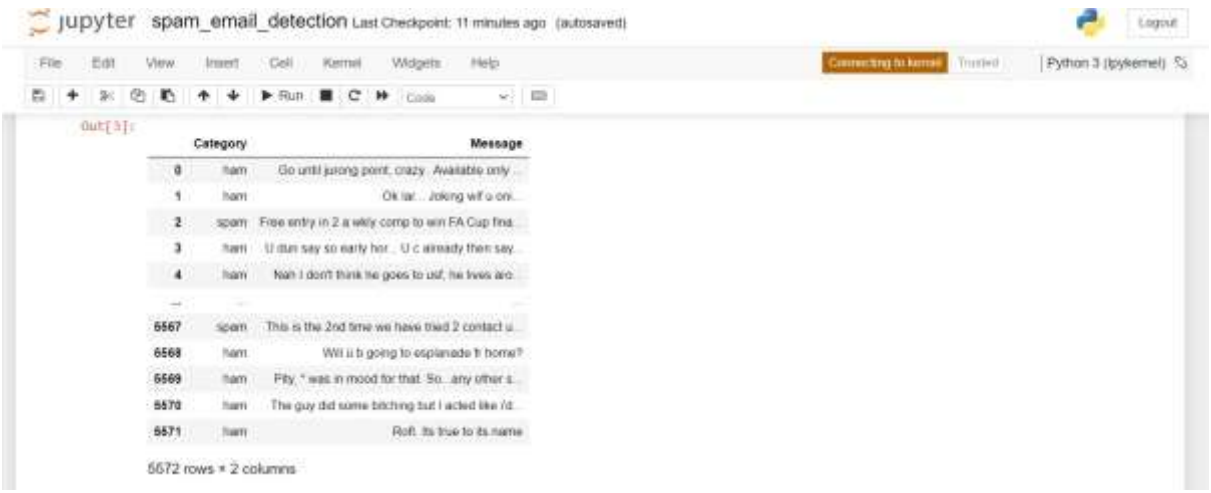

Fig 2. Spam Data

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Fig 3. Total Spam and Ham data set

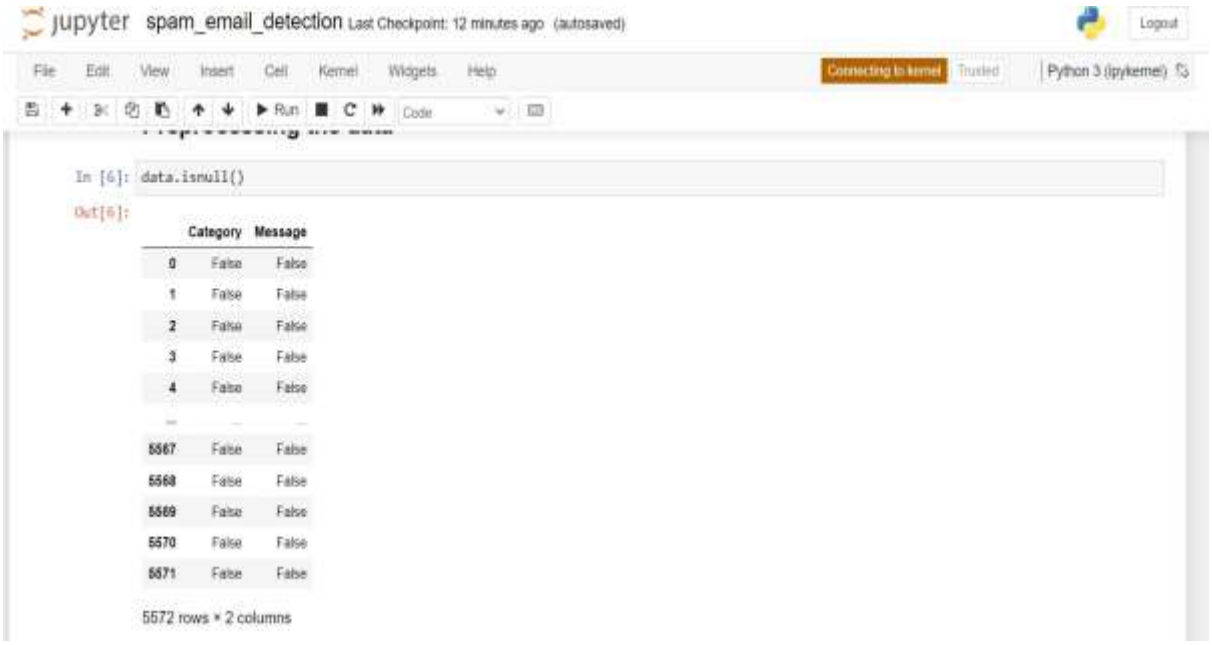

Fig 4. After the removing the null values in Spam data set

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

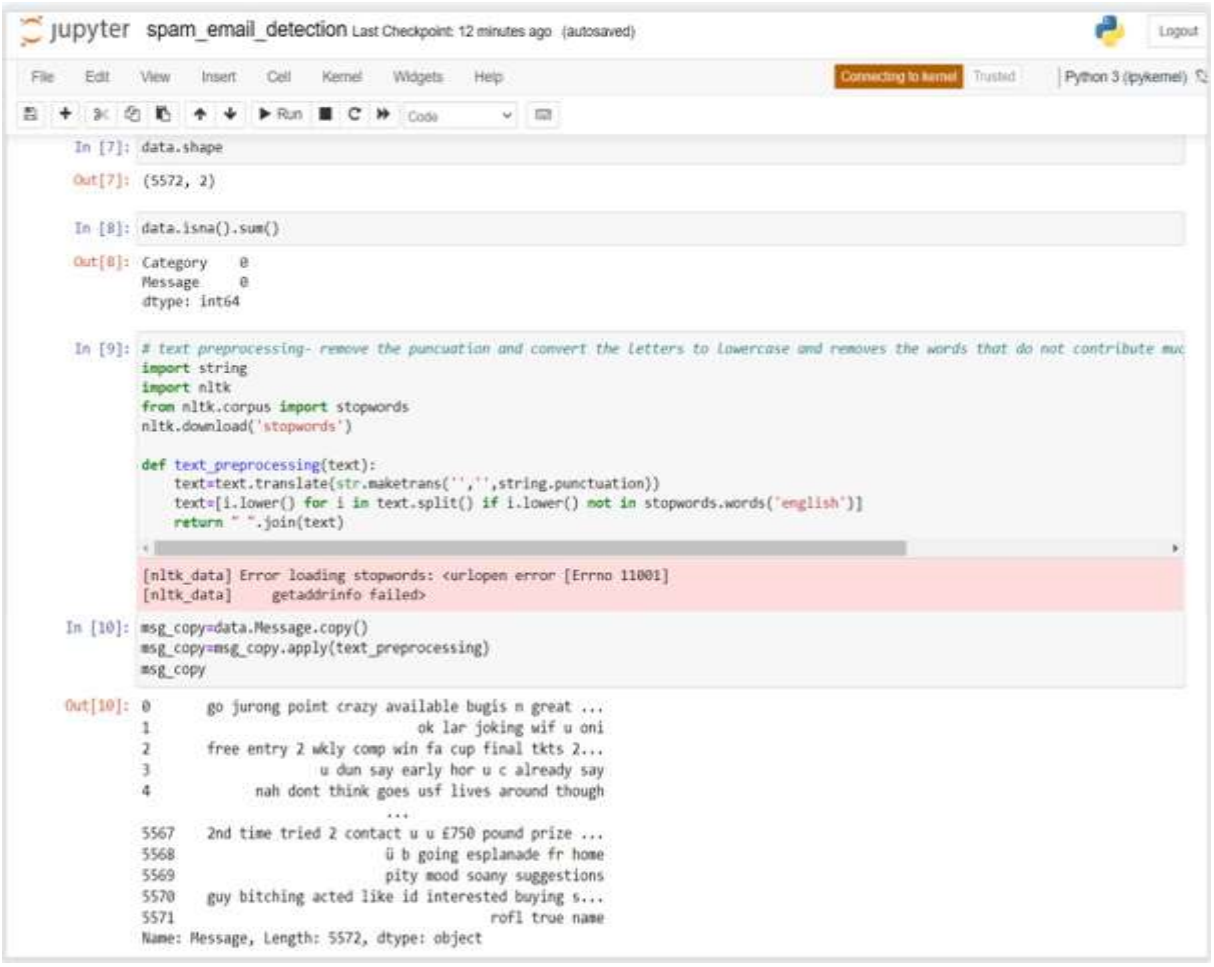

Fig 5. Text Processing

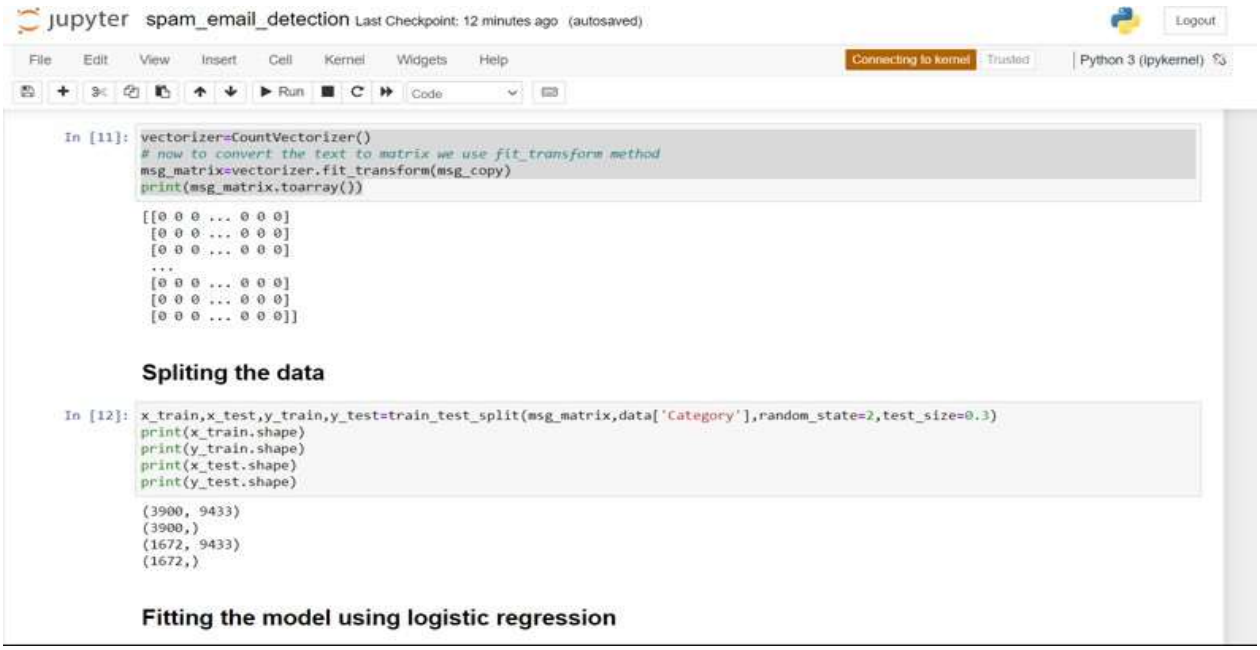

Fig 6. Split data

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

After fitting the model using Naive Bayes Classifier we will compare the performance of both the classifiers.

The classification report, confusion matrix ,Accuracy score, heat map on prediction are measured to calculate the performance of Naive Bayes and Logistic regression. Accuracy is used to determine the performance of Naive Bayes Classifier and Logistic Regression. Using the spam data set, it is necessary to compare and analyse the models. The results clearly show that the Naive Bayes Classifier is more accurate than Logistic Regression. The mean accuracy of the Naive Bayes Classifier is 97.30%, whereas the mean accuracy of the Logistic Regression is 96.77%.

Table 1 shows the statistical variables for Naive Bayes and Logistic Regression that are measured, including precision, recall, f1-score, and support.

Compared to the accuracy of Logistic Regression, the accuracy is 97.30% higher [9].

Table 2 shows the statistical variables for Naive Bayes and Logistic Regression that are measured, including precision, recall, f1-score, and support.

Compared to Naive Bayes, the accuracy is 96.77% which is lower.

Naïve Bayes Model

Evaluating test data(20% of dataset)

	precision	recall	f1-score	support
ham	0.99	0.98	0.98	1445
spam	0.89	0.91	0.90	227
accuracy			0.97	1672
macro avg	0.94	0.95	0.94	1672
weighted avg	0.97	0.97	0.97	1672

confuion matrix:
[[1420 25]
[20 207]]

Accuracy score: 0.9730861244019139

Logistic Regression Model

Evaluating test data(20% of dataset)

	precision	recall	f1-score	support
ham	0.97	0.99	0.98	1445
spam	0.95	0.80	0.87	227
accuracy			0.97	1672
macro avg	0.96	0.90	0.93	1672
weighted avg	0.97	0.97	0.97	1672

confuion matrix:
[[1436 9]
[45 182]]

Accuracy score: 0.9677033492822966

Table 1Table 2

The total Spam and ham data set is shown in Fig 8

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

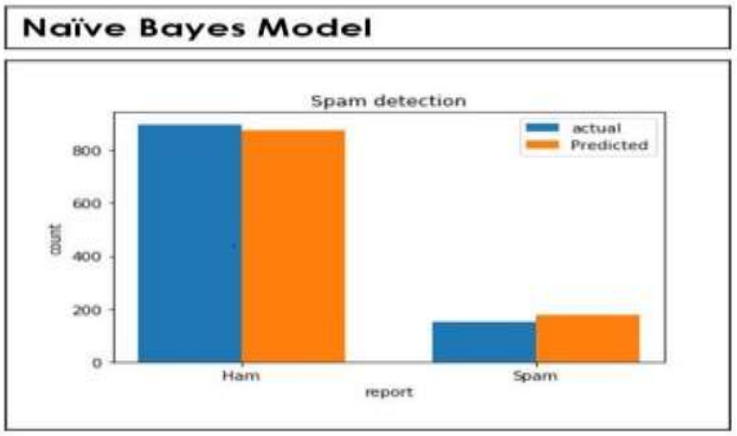

Fig 8. Total spam and ham in the dataset

Evaluation metrics are utilized for gauging the effectiveness of a machine learning model or algorithm. They play a crucial role in determining the model's performance and whether it meets the intended goals. Fig 9 and 10 the Heat Map of predictions are shown below.

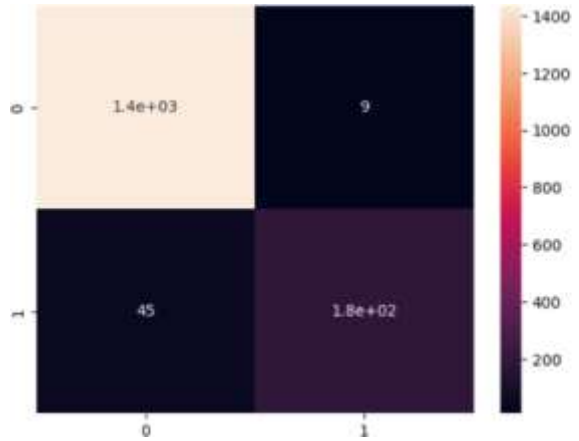

Fig 9. Performance of logistic regression

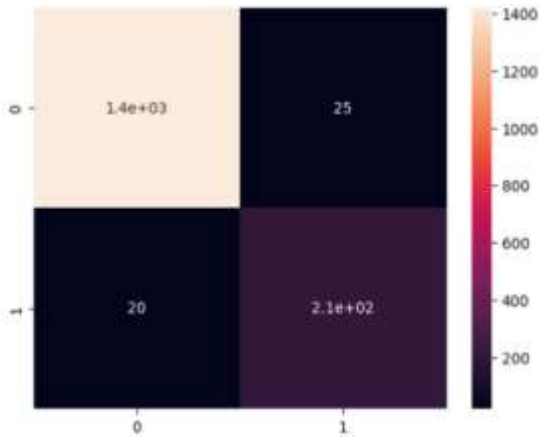

Fig 10. Performance of Naive Bayes

5. Discussion

We plan to create an API for the same in the future and evaluate it in a practical environment. This project will be optimized for datasets to the greatest extent possible.Since the addition successfully enhances the Naive Bayes spam filter.Naive Bayes is aprobabilistic algorithm that determines whether an email is likely to be spam or
legitimate using Bayes' theorem. Naive Bayes makes the unfounded assumption that thefeatures (words) in an email are independent of one another. Naive Bayes has been demonstrated to perform well in practise and is frequently used in spam detection, despite this presumption.

A logistic function is used in the classification algorithm known as logistic regression todetermine whether an email is likely to be spam or legitimate. Logistic regression does not rely on the assumption of feature independence like Naive Bayes does. Non-linear relationships between the features and the target variable can also be handled by logisticregression [10] . A powerful spam detector for text modifications will eventually be created by combining these additional techniques, allowing spam detection to be
improved across a variety of systems.We will try to implement with other deep learningalgorithms.

6. Conclusion

Email can be categorized as spam by adding something to the Naive Bayes Classifier. The high recall and precision rates of our new addition were also found to contribute to an improvement in ham classification. We showed that our algorithm consistently decreased the number of spam emails that were mistakenly labeled as ham emails. The review demonstrates that when compared to the Logistic Regression model, the Naive Bayes classifier has higher accuracy in identifying spam. The Naive Bayes Classifier has a 97.30% accuracy score, which is higher than the Logistic Regression model's 96.77% rating.

7. References

[1]. GitHub, Inc, "Spam Assassin," 21 April 2016. [Online]. Available: https://github.com/dmitrynogin/SpamAssassin.git. [Accessed 20 August 2017].

[2]. Team, Radicati. "Email Statistics Report, 2015-2019. The Radicati Group."(2015).

[3]. Wang, W. B., F. Yin, H. Sun, and P. Li. 2015. “Random Forest Algorithm for Spam Filtering Based on Machine Learning.” In *Electronic Engineering and Information Science*,225–28. CRC Press.

[4]. A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab that uses machinelearning techniques for email spam detection.

[5]. Harisinghaney based on hybrid approach of TF-IDF (Term Frequency InverseDocument Frequency) and is used by Mohamad & Selamat (2015).

[6]. Agarwal, Kriti, and Tarun Kumar. 2018. “Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization.” *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iccons.2018.8662957.

[7]. Singh, Manmohan, Rajendra Pamula, and Shudhanshu Kumar Shekhar. 2018. “EmailSpam Classification by Support Vector Machine.”

[8]. Intelligent Model for Classification of SPAM and HAM.

[9]. Wei, Qijia. 2018. “Understanding of the Naive Bayes Classifier in Spam Filtering.” https://doi.org/10.1063/1.5038979.

[10]. Trivedi, Shrawan Kumar. 2016. “A Study of Machine Learning Classifiers for Spam Detection.” *2016 4th International Symposium on Computational and BusinessIntelligence (ISCBI). https://doi.org/10.1109/iscbi.2016.7743279.*